(12) UK Patent Application (19) GB (11) 2 369 899 (13) A

(43) Date of A Publication 12.06.2002

(21) Application No 0017740.2

(22) Date of Filing 20.07.2000

(51) INT CL[7]
G06K 9/62

(52) UK CL (Edition T)
G4A AUXP

(56) Documents Cited
WO 00/28473 A1

(58) Field of Search
ONLINE: EPODOC, WPI, PAJ, INTERNET

(71) Applicant(s)
Volodya Vovk
Dept of Computer Science,
Royal Holloway University of London, EGHAM,
Surrey, TW20 0EX, United Kingdom

Alex Gammerman
Dept of Computer Science,
Royal Holloway University of London, EGHAM,
Surrey, TW20 0EX, United Kingdom

Royal Holloway University of London
(Incorporated in the United Kingdom)
EGHAM, Surrey, TW20 0EX, United Kingdom

(72) Inventor(s)
Volodya Vovk
Alex Gammerman

(74) Agent and/or Address for Service
Volodya Vovk
Dept of Computer Science, Royal Holloway
University of London, EGHAM, Surrey, TW20 0EX,
United Kingdom

(54) Abstract Title
Data labelling device and method thereof

(57) The present invention relates to data labelling apparatus and to a method thereof that is capable of identifying for an unknown example a range of most suitable labels and that is additionally able to provide a measure of confidence, which is valid under the general iid assumption, in the range identified; a priori there may be a large number, often an infinite range, of potential labels. A typical drawback of currently used data labelling apparatuses is that the user is not provided with any measure of the accuracy of the predicted output by the apparatus; in cases where such a measure is given, it is only valid under strong extra assumptions. The present invention thus seeks to provide apparatus and a method to identify potential labels for an unlabelled example and that is able to generate a valid and practicable measure of confidence for the potential labels identified.

GB 2 369 899 A

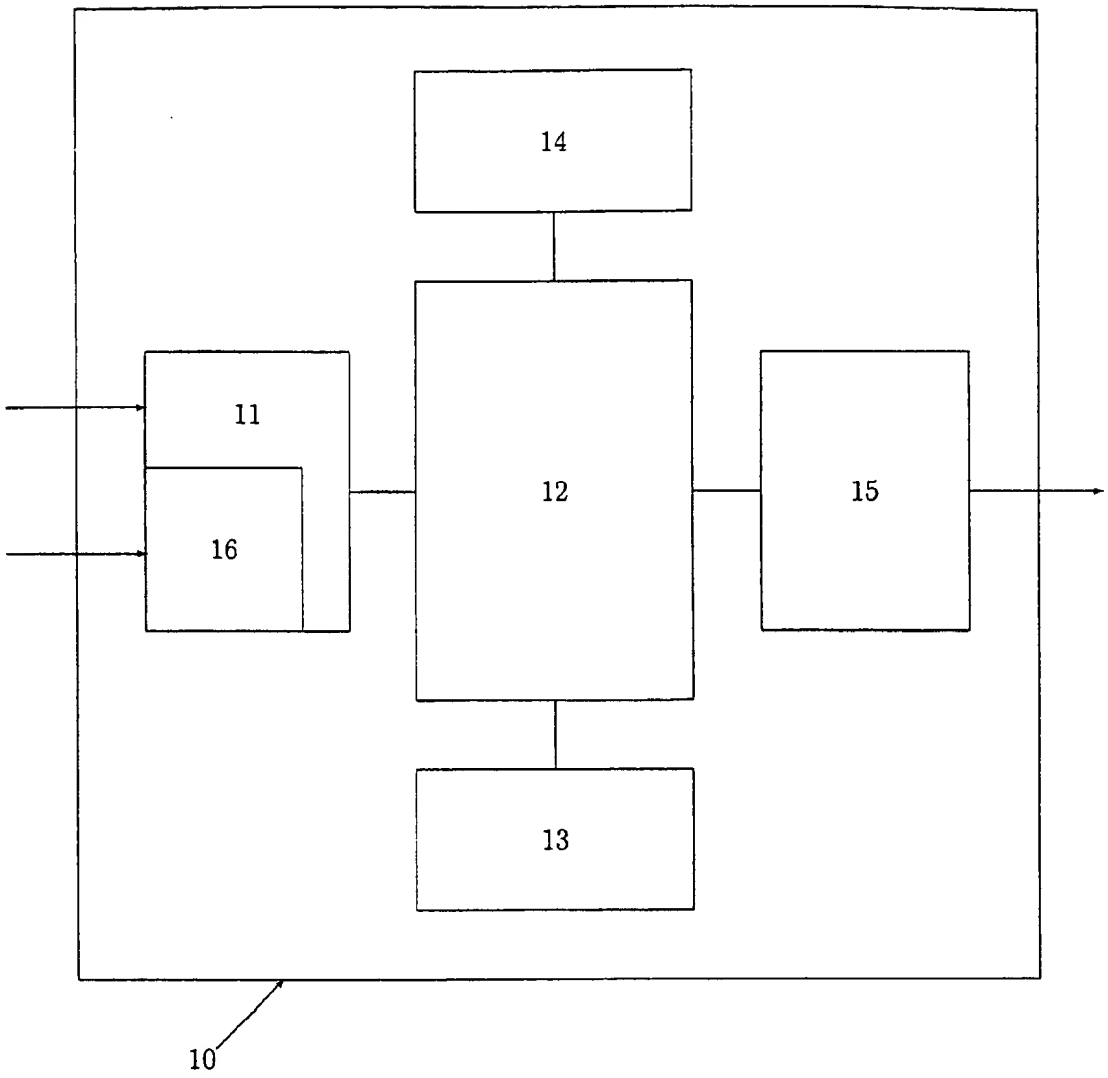

Figure 1

Training set

Example No	x	y
1	0	0
2	1	0.5
3	0.5	3
4	2	0.8

Test set

Example No	x	y
5	0.7	
6	-17	

Figure 2

Training set (with 7 attributes and label ECC/t)

Systolic BP	Diastolic BP	Sex	Mean QP	Mean cholesterol	ACE I	DG	ECC/t
166	79	1	3.91	9.03	1	7	−1.07
165	100	1	1.5	5.57	2	4	−0.10
148	69	2	0.95	5.9	2	1	−0.39
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
162	75	1	2.3	3.8	2	6	−0.45
141	79	1	3.97	4.55	1	7	−0.67

Test set

Systolic BP	Diastolic BP	Sex	Mean QP	Mean cholesterol	ACE I	DG	ECC/t
136	92	1	0.76	5.83	1	3	−0.12
151	98	1	1.7	7.15	2	1	−0.60

Figure 3

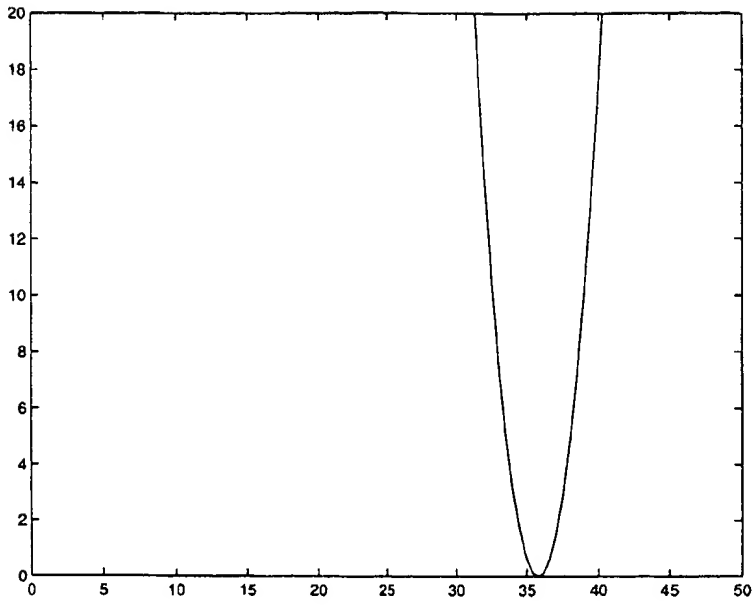

Figure 4

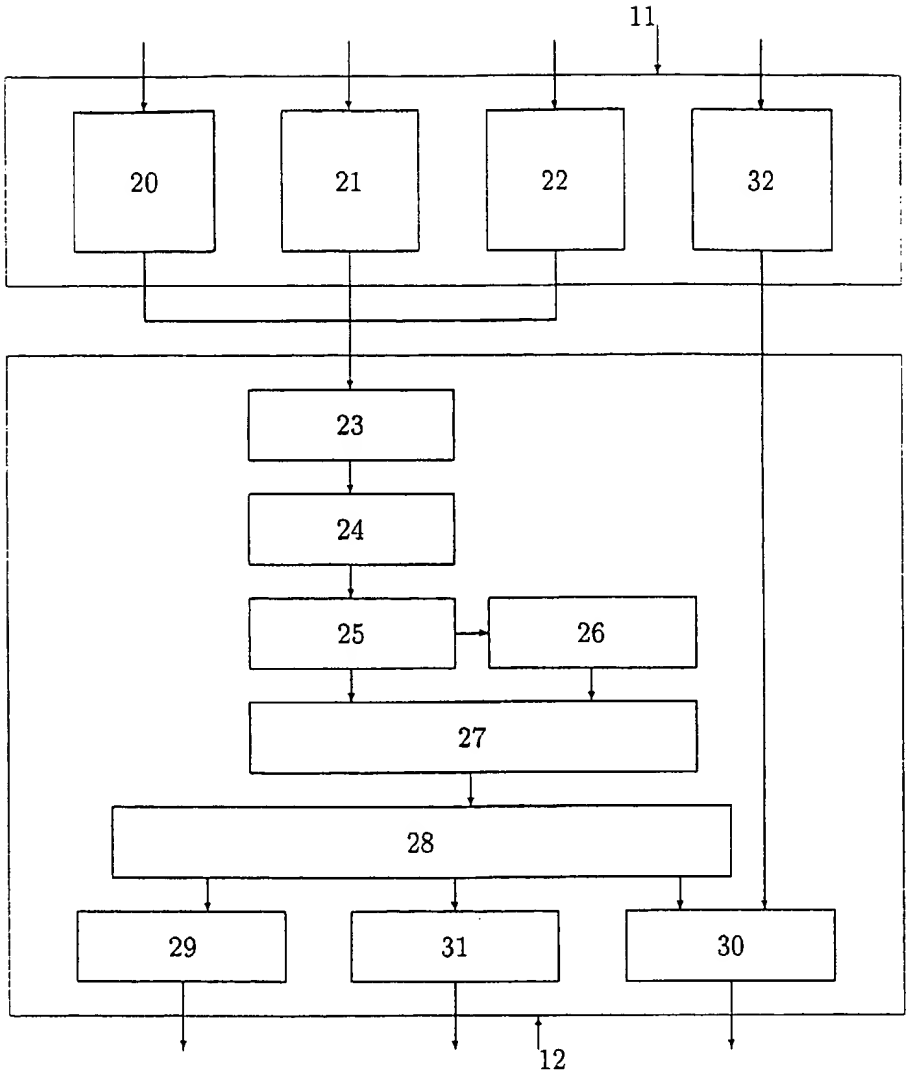

Figure 5

DATA LABELLING APPARATUS AND METHOD THEREOF

The present invention relates to data labelling apparatus and to a method thereof that is capable of identifying for an unknown example a range of most suitable labels and that is additionally able to provide a measure of confidence in the range identified.

In the context of this document it is to be understood that data labelling is intended as reference to the labelling of new, unlabelled, examples for which there is a large number, often an infinite range, of potential labels. This is in contrast to data classification, which is usually concerned with a very limited number, often only two, potential classifications.

A practical example of data labelling is in the assessment of house values. The range of possible values for the building is infinite. In practice, the actual range of likely values is much smaller and is dependent on such factors as number of bedrooms, location, state of repair etc. Using the data labelling technique described herein a range of potential values for an individual house can be generated automatically avoiding the subjective assessment usually involved in such valuations. Another practical example is in optimising the operating characteristics of a complex on-line manufacturing process.

Learning machines that have already been developed to perform data labelling include Support Vector machines (described in V. N. Vapnik, Statistical Learning Theory, New York: Wiley, 1998) and Ridge Regression machines. A paper describing a learning machine employing Ridge Regression in data labelling may be found in Machine Learning, Proceedings of the Fifteenth International Conference, pp. 515–521, entitled "Ridge Regression Learning Algorithm in Dual Variables", C. Saunders, A. Gammerman and V. Vovk. Some of these known machines perform very well in a wide range of applications and do not require any parametric statistical assumptions about the source of the data (unlike traditional statistical procedures); the only assumption is that the examples are generated from the same distribution independently of one another—the i.i.d. assumption.

A typical drawback of such machines is that the user is not provided with any measure of the accuracy of the predicted output by the learning machine. A user has to rely on the results of previous experiments with benchmark datasets, with the hope that for the user's particular dataset similar results will be obtained. Other options for the user who wants to associate a measure

of accuracy with new unlabelled examples include performing experiments on a validation set, using one of the known cross-validation procedures, and applying one of the theoretical results, which are usually very crude, about the future performance of different learning machines given their past performance. None of the known accuracy estimation procedures provide any practicable means for directly assessing the accuracy of a predicted "real-world" label for an individual new example in practical machine-learning problems.

Interval estimation, which addresses the problem of accuracy in a rigorous way, is a well-studied area of both parametric and non-parametric statistics. Typically, in statistics one is interested in intervals containing the true values of the parameter (or some component of the parameter in the semi-parametric setting). In traditional statistics, however, no closed-form formulas are derived in the general non-parametric case and only low-dimensional problems can be dealt with.

The present invention thus seeks to provide apparatus and a method that relies upon the Ridge Regression or another conventional technique to identify potential labels for an unlabelled example and that is able to generate a valid measure of confidence for the potential labels identified.

The present invention provides data labelling apparatus comprising:

- an input device for receiving a plurality of training labelled examples and at least one unlabelled example;
- a memory for storing the labelled and unlabelled examples;
- an output terminal for outputting the one or more predicted labels for the at least one unlabelled example; and
- a processor for identifying the one or more predicted labels of the one or more unlabelled example,

wherein the processor includes a program memory in which is stored programming for performing, analytically or computationally, the following steps:

- associating respective individual strangeness values with all or some examples in a plurality of label sets, each label set consisting of the labelled examples and their labels and the at least one unlabelled example with a potential label, the individual strangeness values being defined by means of an optimisation algorithm;

- associating a strangeness value with each label set based on the individual strangeness value for the at least one unlabelled example;
- determining the relationship between potential labels for each unlabelled example and their associated strangeness values; and
- identifying from the relationship one or more predicted labels for the at least one unlabelled example.

With the present invention in addition to a predicted label or range of labels for every unlabelled example, a strangeness value for every possible label is also generated. This strangeness value has a clear interpretation, either as an i-value or as a p-value, in terms of the mathematical theory of probability and is valid under the general i.i.d. assumption. Furthermore, the present invention is particularly suited to dealing with high dimensional problems and where there is a very large number, e.g., more than one million, labels.

In a first embodiment the optimisation algorithm stored in the programming memory is a Ridge Regression procedure. In alternative embodiments the optimisation algorithm stored in the program memory may be the Aggregating Algorithm, the Nearest Neighbours Algorithm, etc.

The labelling programming stored in the program memory may include a program for identifying a minimum strangeness value and for identifying the potential label associated with the minimum strangeness value and for outputting the identified potential label as the predicted label. Additionally, the program memory may include threshold programming for identifying a range of strangeness values less than a predetermined strangeness threshold and for outputting the potential labels associated with the identified range of strangeness values as a range of predicted labels in which case the input may include means for inputting a chosen strangeness threshold. In a further alternative the program memory may include programming for plotting a graphical representation of the relationship of strangeness values with respect to potential labels.

Ideally, the program memory includes one or more programs for transforming the optimisation algorithm using Lagrange multipliers and the program memory may include programming for applying the optimisation algorithm to images of the attribute vectors in a Hilbert space.

In a second aspect, the present invention provides a data labelling method

comprising the following steps that are performed analytically or computationally:

- inputting a plurality of training labelled examples and at least one unlabelled example;
- associating respective individual strangeness values with all or some examples in a plurality of label sets, each label set consisting of the labelled examples and their labels and the at least one unlabelled example with its potential label, the individual strangeness values being defined by means of an optimisation algorithm;
- associating a strangeness value with each label set based on the individual strangeness value for the at least one unlabelled example;
- determining the relationship between potential labels for each unlabelled example and their associated strangeness values;
- identifying from the relationship one or more predicted labels for the at least one unlabelled example; and
- outputting one or more predicted labels for the at least one unlabelled example.

An embodiment of the present invention will now be described by way of example with reference to the accompanying drawings, in which:

- Figure 1 is a schematic diagram of data labelling apparatus in accordance with the present invention;
- Figure 2 is an example of a training set and a test set for use with the present invention;
- Figure 3 is a second example of a training set and a test set for use with the present invention;
- Figure 4 is a plot of a confidence graph;
- Figure 5 is a schematic diagram of a data labelling method in accordance with the present invention.

In Figure 1 a data labeller 10 is shown generally consisting of an input device 11, a processor 12, a memory 13, a ROM 14 containing a suite of programs accessible by the processor 12 and an output terminal 15. The input device 11 preferably includes a user interface 16 such as a keyboard or other conventional means for communicating with and inputting data to the processor 12, and the output terminal 15 may be in the form of a display monitor or other conventional means for displaying information to a user. The output terminal 15 preferably includes one or more output ports for connection to a printer or other network device. The processor 12 and memories 13, 14 may be embodied in an Application Specific Integrated Circuit (ASIC) with additional RAM chips. Ideally the ASIC would contain a fast RISC CPU with an appropriate Floating Point Unit.

To assist in an understanding of the operation of the data labeller 10 in providing a prediction of labels for unlabelled (unknown) examples, the following is an explanation of the mathematical theory underlying its operation.

Two sets of examples (data vectors) are given: the training set that consists of examples with their labels known and a test set that consists of unlabelled examples. Therefore, each example in the training set contains an attribute vector and a label, whereas each example in the test set is identical with an attribute vector. Figures 2 and 3 each exemplify separate training sets and test sets. The size of the training set is given by T and for the sake of simplicity the test set is limited to one unlabelled example. Let X be the set of all possible attribute vectors (e.g., in the case of Figure 3, X might be the Cartesian product $\Re^7$); it is assumed that the set of all possible labels is $\Re$, the real line.

The training set consists of labelled examples $((x_1, y_1), \ldots, (x_T, y_T))$, where T is the number of training examples, x_t are attribute vectors in $\Re^n$ (n being the number of attributes) and $y_t \in \Re$, $t = 1, \ldots, T$. The goal is to predict the label y_{T+1} of the new unlabelled example x_{T+1}.

An important feature of the data labeller is the determination of strangeness values. Although the use of strangeness values is known in algorithmic information theory with respect to the deficiency of randomness, see for example "An introduction to Kolmogorov Complexity and Its Applications", M. Li and P. Vitanyi, strangeness values have not previously been employed in the mathematical field of classification and labelling. The two main types of the deficiency of randomness are those proposed by Per Martin-Löf described in [*Information and Control*, 9:602–619, 1966] and by

Leonid Levin [described in, e.g., "On the Empirical Validity of the Bayesian Method" by V. Vovk and V. V'yugin, *J. R. Statist. Soc.* B, 55:253–266, 1993]. However, neither of these two types is computable; an approximation has therefore been developed that is computable. The approximation is based on the notions of a randomness test and a measure of impossibility, as discussed in the papers referred to above.

In order to develop a mathematical basis for the measure of impossibility, let Ω be a sample space (a typical sample space is the set $(X \times \Re)^{T+1}$ of all label sets, i.e., sequences $(x_1, \ldots, x_{T+1})$ of $T+1$ points in the Euclidean space $x_t \in \Re^n$ with their labels $y_t \in \Re$, $t = 1, \ldots, T+1$). If P is a probability distribution in Ω, a P-measure of impossibility is defined to be a non-negative measurable function $p : \Omega \to \Re$ such that

$$\int_\Omega p(\omega) P(d\omega) \le 1. \tag{1}$$

This provides a notion of a "lottery" in which P is a randomising device used for drawing lots and $p(\omega)$ is the value of the prize won by a particular ticket when P produces ω. With equation (1) "fair" lotteries, in which equation (1) is satisfied with an equality sign, (i.e., lotteries in which all proceeds from selling the tickets are redistributed in the form of prizes) are not excluded. In reality, for lotteries the left-hand side of equation (1) is usually much less than 1.

By Chebyshev's inequality, p is large with small probability: for any constant $C > 0$,

$$P\{\omega \in \Omega : p(\omega) \ge C\} \le \frac{1}{C}.$$

This confirms that if p is chosen in advance and P is assumed to be the true probability distribution generating the data $\omega \in \Omega$, then it is unlikely $p(\omega)$ will turn out to be large. Hence, $p(\omega)$ is taken to be the strangeness value assigned to ω by p. Its inverse $1/p(\omega)$ is called the i-value assigned to ω.

The above, though, is concerned with a single distribution P. If μ is a family of probability distributions, a μ-measure of impossibility is defined as a function which is a P-measure of impossibility for all $P \in \mu$. For the purposes of data labelling, the $\mathcal{P}^m(Z)$-measure of impossibility is of interest where Z is any measurable space, m is a positive integer (the sample size) and $\mathcal{P}^m(Z)$ stands for the set of all product distributions P^m in Z^m, P running over all probability distributions in Z. This definition is interpreted as follows: if p is a $\mathcal{P}^m(Z)$-measure of impossibility and $z_1, \ldots, z_m$ are generated independently

from the same distribution (the i.i.d. assumption), it is hardly possible that $p(z_1, \ldots, z_m$ is large (provided p is chosen before the data $z_1, \ldots, z_m$ are generated).

In data labelling m (the sample size) equals $T+1$ and Z (the measurable space) equals $X \times \Re$ such that $\mathcal{P}^{T+1}(X \times \Re)$-measures of impossibility are of interest.

In order to determine a particular $\mathcal{P}^{T+1}(X \times \Re)$-measure of impossibility, a continuum of completions is considered of the available data: $(x_1, y_1), \ldots, (x_T, y_T), x_{T+1}$. The completion y where $y \in Y$ is $(x_1, y_1), \ldots, (x_T, y_T), (x_{T+1}, y)$ (thus in all completions every example is labelled); such completions will be called label sets. In the following explanation y is temporarily denoted as y_{T+1} for the sake of clarity. Some strangeness value must be associated with each label set $(x_1, y_1), \ldots, (x_{T+1}, y_{T+1})$. This is done by defining individual strangeness values in terms of an auxiliary optimisation problem.

For example, with every label set $(x_1, y_1), \ldots, (x_T, y_T), (x_{T+1}, y)$ is associated a Ridge Regression optimisation problem

$$a(w \cdot w) + \sum_{t=1}^{T+1} (y_t - w \cdot x_t)^2 \to \min, \tag{2}$$

where $a > 0$ is a fixed constant. There is an implicit assumption here that some linear function $x \mapsto y$ fits the data well; later this assumption is dispensed with. The above problem is then rewritten introducing slack variables ξ_t as

$$a(w \cdot w) + \left(\sum_{t=1}^{T+1} \xi_t^2 \right) \to \min, \tag{3}$$

subject to the constraints

$$\xi_t = y_t - ((x_t \cdot w) + b), \ t = 1, \ldots, T+1. \tag{4}$$

As usual in the art, this optimisation problem is transformed, via the introduction of Lagrange multipliers α_t, $t = 1, \ldots, T+1$ to the dual problem: find α_t from

$$\sum_{t=1}^{T+1} y_t \alpha_t - \frac{1}{4} \sum_{t=1}^{T+1} \alpha_t^2 - \frac{1}{4a} \frac{1}{2} \sum_{t,s=1}^{T+1} y_t y_s \alpha_t \alpha_s (x_t \cdot x_s) \to \max. \tag{5}$$

This particular optimisation problem can be solved explicitly providing the solution

$$\hat{y} = Y'(K + aI)^{-1}k. \tag{6}$$

In equation (6) the following notation is employed: Y is the vector of the first T labels,

$$\begin{pmatrix} y_1 \\ \vdots \\ y_T \end{pmatrix},$$

K is the $T \times T$ matrix from $x_1, \ldots, x_T$,

$$K_{t,s} = x_t \cdot x_s, \;\; t = 1, \ldots, T, \;\; s = 1, \ldots, T,$$

and k is the vector

$$\begin{pmatrix} x_1 \cdot x_{T+1} \\ \vdots \\ x_T \cdot x_{T+1} \end{pmatrix}.$$

The square α_t^2 of the Lagrange multiplier α_t is taken as the individual strangeness value of (x_t, y_t). This is proportional to the squared distance (measured along the y-axis) from (x_t, y_t) to the best Ridge Regression approximation to the label set $(x_1, y_1, \ldots, x_{T+1}, y_{T+1})$. The measure of impossibility of the label set will be defined as the individual strangeness value, properly normalised, of the last example (x_{T+1}, y_{T+1}), thus as the measure of impossibility the following ratio is used:

$$\frac{\alpha_{T+1}^2}{\frac{1}{T+1}\sum_{t=1}^{T+1}\alpha_t^2}.$$

This results in the measure of impossibility being rewritten as:

$$\begin{aligned}(T+1)(y-\hat{y})^2 / \Big(&\|(K+aI)^{-1}Y(\|x_{T+1}\|^2 + a - k'(K+aI)^{-1}k) \\ &+(K+aI)^{-1}k(\hat{y}-y)\|^2 + (y-\hat{y})^2 \Big),\end{aligned} \tag{7}$$

where $\hat{y}$ is the Ridge Regression prediction in equation (6) of y_{T+1}. Thus, where $y \approx \hat{y}$, the measure of impossibility is low whereas where y is very different from $\hat{y}$ the measure of impossibility is high.

Evaluation of equation (7) can be implemented as follows:

- Compute matrix $B = (K + aI)^{-1}$
- Compute vector $V = Bk$
- Compute vector $U = BY(\|x_{T+1}\|^2 + a - k'V)$
- Compute numbers $\|U\|^2$, $U \cdot V$ and $\|V\|^2$
- Plot (as a function of $z = y - \hat{y}$) the confidence graph

$$(T+1)\frac{z^2}{\|U - Vz\|^2 + z^2} = \frac{(T+1)z^2}{\|U\|^2 - 2(U \cdot V)z + (\|V\|^2 + 1)z^2} \qquad (8)$$

An example of such a plot is shown in Figure 4.

A typical mode of use of this formula is that some threshold, such as 20 or 100, is chosen in advance; e.g., choosing 20 means that we regard winning £20 or more on a £1 lottery ticket unlikely. (This corresponds to choosing one of the standard significance levels such as 5% or 1% in statistics). After this the prediction might be the smallest interval containing labels with strangeness values at most 20.

Next the linearity assumption is removed. The quadratic optimisation problem, equation (2), is applied not to the attribute vectors x_t themselves, but to their images $F(x_t)$ under some predetermined function $F : X \to H$ taking values in a Hilbert space, which leads to replacing the dot product $x_t \cdot x_s$ in the optimisation problem in equation (5) by the kernel function

$$\kappa(x_t, x_s) = F(x_t) \cdot F(x_s).$$

The final expression for the confidence graph is, therefore, (7) with K and k defined using the kernel function, i.e., K defined to be the matrix

$$K_{t,s} = \kappa(x_t, x_s), \quad t = 1, \ldots, T, \quad s = 1, \ldots, T,$$

an k the vector

$$\begin{pmatrix} \kappa(x_1, x_{T+1}) \\ \vdots \\ \kappa(x_T, x_{T+1}) \end{pmatrix}$$

With the data labelling apparatus of the present invention the following menus or choices may be offered to a user:

1. Prediction

2. Prediction with a given threshold for the measure of impossibility

3. Complete plot of the confidence graph

A typical response to the user's selection of choice 1 might be "Prediction: 36", which means 36 will be the predicted output. A typical response to the selection of choice 2 might be "Predictive interval: [32,40]", which gives the smallest interval containing the labels whose strangeness value does not exceed the chosen threshold (such as 20). A typical response to the selection of choice 3 might be the confidence graph of Figure 4 which is the complete plot of the strangeness values of all potential labels. It will be apparent that the "prediction" of choice 1 is where the minimum of the plot is obtained.

It is contemplated that some modifications of the optimisation problem set out in equations (3) and (4) might have certain advantages, for example the Support Vector problem:

$$a(w \cdot w) + \left(\sum_{t=1}^{T+1} \xi_t \right) \to \min,$$

subject to the constraints

$$|y_t - ((x_t \cdot w) + b)| \leq \epsilon + \xi_t,\ \xi_t \geq 0, \quad t = 1, \ldots, T+1.$$

An alternative optimisation problem (for which a closed-form formula can be easily derived) that may be employed is provided by the Aggregating Algorithm as described in "Competitive on-line linear regression", V. Vovk in *Advances in Neural Information Processing Systems*, pages 364–370, Cambridge MA, 1998.

It is further contemplated that the data labelling apparatus will be particularly useful for predicting the labels of more than one unlabelled example using a closed-form formula for computing the strangeness values corresponding to different completions. These strangeness values can be provided not only by measures of impossibility, but also by randomness tests, which would correspond to using the statistical notion of p-values in place of i-values.

In practice, as shown in Figure 5, a training dataset is input 20 to the data labeller. The training dataset consists of a plurality of data vectors $(x_1, \ldots, x_T)$ each of which has an associated known label $(y_1, \ldots, y_T)$ allocated. Some constructive representation of the measurable space of the data

vectors is input 21 to the data labeller or stored in the ROM 14. For example, in the case of Figure 3, the measurable space might be $\Re^7$ or in the case of house prices the measurable space might consist of the number of rooms, the size of any garden, garaging and location etc. Where the measurable space is already stored in the ROM 14 of the data labeller, the interface 16 may include input means (not shown) to enable a user to input adjustments for the stored measurable space. For example, a more precise definition of a location by street or area may be needed.

One or more data vectors (x_{T+1}) for which no label is known are also input 22 into the data labeller. The training dataset and the unlabelled data vectors along with any additional information input by the user are then fed from the input device 11 to the processor 12.

Label sets are then identified containing each of the labelled examples with their labels and the unlabelled examples with their provisional labels. Associated individual strangeness values are then defined by means of an optimisation algorithm such as the Ridge Regression procedure. Strangeness values are then defined for the unclassified examples from the individual strangeness values. The relationship between potential labels for each unlabelled example and their associated strangeness values is then determined and from the relationship one or more predicted labels for each unlabelled example is identified.

To do this using the Ridge Regression optimisation problem, the matrix K of the kernel function (which replaces the dot product $(x_t \cdot x_s)$) is determined 23. Next the matrix B is determined 24 from $B = (K + aI)^{-1}$ and then the vector V is determined 25 from $V = Bk$, where k is the vector of the product of each training attribute vector with the unlabelled attribute vector. The vector U is also determined 26 using the matrix B and vector V and then values of $\|U\|^2$, $U \cdot V$ and $\|V\|^2$ are calculated 27. Finally equation (7) is used to determine a confidence graph 28 of the measure of impossibility for the potential labels of the unlabelled data vector x_{T+1}. The minimum of the confidence graph is output 29 as the prediction for choice 1, a range of labels having less than a predetermined (or supplied 32 by the user) impossibility threshold is output 30 in response to choice 2 and a plot of the entire confidence graph is output 31 in response to choice 3. Preferably, the predetermined threshold may be stored in the ROM 14.

Although the above description of the data labelling apparatus and method uses the example of assigning values to houses it is to be understood that the data labelling apparatus and method may be used in a wide

variety of useful applications, for example: the time to failure of a mechanical component. Further examples might be estimating a patient's level of renal decline before taking more expensive tests (the figures given in Figure 3 relate to renal decline), or estimating the target company's future profits before a take-over. It is clear that confidence measures are very useful in such applications (especially in safety-critical situations): e.g., a decision might be made to arrange for more expensive tests even for a patient with low estimated renal decline if the confidence in the estimate of renal decline is low.

While the data labelling apparatus and method described above has been particularly shown and described with reference to the preferred embodiment, it will be understood by those skilled in the art that various modifications in form and detail may be made therein without departing from the scope and spirit of the invention. Accordingly, modifications such as those suggested above, but not limited thereto, are to be considered within the scope of this invention.

CLAIMS

1. Data labelling apparatus comprising:

   - an input device for receiving a plurality of training labelled examples and at least one unlabelled example;
   - a memory for storing the labelled and unlabelled examples;
   - an output terminal for outputting one or more predicted labels for the at least one unlabelled example; and
   - a processor for identifying the one or more predicted labels of the at least one unlabelled example,

   wherein the processor includes a program memory in which is stored programming for performing analytically or computationally the following steps:

   - associating respective individual strangeness values with all or some examples in a plurality of label sets, each label set consisting of the labelled examples and their labels and the at least one unlabelled example with a potential label, the individual strangeness values being defined by means of an optimisation algorithm;
   - associating a strangeness value with each label set based on the individual strangeness value for the at least one unlabelled example;
   - determining the relationship between potential labels for each unlabelled example and their associated strangeness values; and
   - identifying from the relationship one or more predicted labels for the at least one unlabelled example.

2. Data labelling apparatus as claimed in claim 1, wherein the optimisation algorithm stored in the program memory is the Ridge Regression algorithm.

3. Data labelling apparatus as claimed in claim 1, wherein the optimisation algorithm stored in the program memory is a Nearest Neighbours algorithm.

4. Data labelling apparatus as claimed in claim 1, wherein the optimisation algorithm stored in the program memory is the Aggregating Algorithm.

5. Data labelling apparatus as claimed in claim 1, wherein the optimisation algorithm stored in the program memory is the Support Vector Machine.

6. Data labelling apparatus as claimed in claim 1, wherein the optimisation algorithm stored in the program memory is a neural network.

7. Data labelling apparatus as claimed in any one of claims 1 to 6, wherein the program memory includes programming for identifying a range of strangeness values less than a predetermined strangeness threshold and for outputting the potential labels associated with the identified range of strangeness values as a range of predicted labels.

8. Data labelling apparatus as claimed in claim 7, wherein the input device includes means for inputting a chosen strangeness threshold.

9. Data labelling apparatus as claimed in any one of claims 1 to 6, wherein the program memory includes programming for outputting a graphical representation of the relationship of strangeness values with respect to potential labels.

10. Data labelling apparatus as claimed in any one of the preceding claims, wherein the program memory includes programming for transforming the optimisation algorithm using Lagrange multipliers.

11. Data labelling apparatus as claimed in any one of the preceding claims, wherein the program memory includes programming for applying the optimisation algorithm to images of the attribute vectors in a Hilbert space.

12. Data labelling apparatus as claimed in any one of the preceding claims, wherein part of the training set is dedicated as a calibration set, so that the strangeness value for a label set depends only on the individual strangeness values for the test examples and the examples in the calibration set.

13. A data labelling method comprising the following steps that are performed analytically or computationally:

    - inputting a plurality of training labelled examples and at least one unlabelled example;
    - associating respective individual strangeness values with all or some examples in a plurality of label sets, each label set consisting of the labelled examples and their labels and the at least one unlabelled example with a potential label, the individual strangeness values being defined by means of an optimisation algorithm;
    - associating a strangeness value with each label set based on the individual strangeness value for the at least one unlabelled example;
    - determining the relationship between potential labels for each unlabelled example and their associated strangeness values;
    - identifying from the relationship one or more predicted labels for the at least one unlabelled example, and
    - outputting the one or more predicted labels for the at least one unlabelled example.

14. A data labelling method as claimed in claim 13, wherein the optimisation algorithm used to define the strangeness values is the Ridge Regression algorithm.

15. A data labelling method as claimed in claim 13, wherein the optimisation algorithm used to define the strangeness values is a Nearest Neighbours algorithm.

16. A data labelling method as claimed in claim 13, wherein the optimisation algorithm used to define the strangeness values is the Aggregating Algorithm.

17. A data labelling method as claimed in claim 13, wherein the optimisation algorithm used to define the strangeness values is the Support Vector Machine.

18. A data labelling method as claimed in claim 13, wherein the optimisation algorithm used to define the strangeness values is a neural network.

19. A data labelling method as claimed in any one of claims 13 to 18, further comprising the steps of identifying a range of strangeness values less than a predetermined threshold and outputting the labels associated with the identified range of strangeness values as a range of predicted labels.

20. A data labelling method as claimed in claim 19, further comprising inputting a chosen strangeness threshold.

21. A data labelling method as claimed in any one of claims 13 to 18, further comprising plotting the relationship of strangeness values with respect to potential labels.

22. A data labelling method as claimed in any one of claims 13 to 21, wherein the optimisation algorithm is transformed using Lagrange multipliers.

23. A data labelling method as claimed in any one of claims 13 to 22, wherein the optimisation algorithm is applied to images of the attribute vectors in a Hilbert space.

24. A data labelling method as claimed in any one of claims 13 to 23, wherein part of the training set is dedicated as a calibration set, so that the strangeness value for a label set depends only on the individual strangeness values for the test examples and the examples in the calibration set.

The Patent Office
PATENTS • DESIGNS • COPYRIGHT • TRADE MARKS
INVESTOR IN PEOPLE

Application No: GB 0017740.2 **Examiner:** Steven Gross
Claims searched: 1-24 **Date of search:** 5 April 2002

17/

Patents Act 1977
Search Report under Section 17

Databases searched:

UK Patent Office collections, including GB, EP, WO & US patent specifications, in:

UK Cl (Ed.T):

Int Cl (Ed.7):

Other: Online: EPODOC, WPI, PAJ, Internet

Documents considered to be relevant:

Category	Identity of document and relevant passage	Relevant to claims
X	WO 00/28473 A1 (ROYAL HOLLOWAY) See whole document	1 & 13 at least

X	Document indicating lack of novelty or inventive step	A	Document indicating technological background and/or state of the art.
Y	Document indicating lack of inventive step if combined with one or more other documents of same category.	P	Document published on or after the declared priority date but before the filing date of this invention.
&	Member of the same patent family	E	Patent document published on or after, but with priority date earlier than, the filing date of this application.